УДК 004.8+004.9

Машечкин И.В., Петровский М.И., Поспелова И.И., Царёв Д.В.

Московский государственный университет им. М.В. Ломоносова, г. Москва, Россия

МЕТОДЫ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ И ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ В ЗАДАЧАХ ОБНАРУЖЕНИЯ ЭКСТРЕМИСТКОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ*

АННОТАЦИЯ

В настоящее время увеличивается число и растет ущерб от террористических атак, осуществляемых как террористами одиночками под воздействием пропаганды и экстремистской идеологии, так и организованными террористическими сообществами, имеющими сетевую, слабо связную структуру. Основным средством обмена информацией, рекрутинга и пропаганды для таких структур является сеть Интернет, а именно веб ресурсы, социальные сети и электронная почта. В связи с этим возникает задача обнаружения, выявления тематик общения, связей, а также мониторинга поведения и прогнозирования угроз, исходящих от отдельных пользователей, групп и сетевых сообществ, порождающих и распространяющих террористическую и экстремистскую информацию в Интернете. Настоящая работа посвящена исследованию и разработке методов машинного обучения, направленных на решение задач обнаружения потенциально опасной информации в сети Интернет. Предложен метод автоматического аннотирования и выявления ключевых слов для поиска информации экстремистского содержания в потоках текстовых сообщений. Экспериментально показана применимость и эффективность предложенного метода на эталонном наборе данных, собранном в рамках проекта Dark Web.

КЛЮЧЕВЫЕ СЛОВА

Безопасность и противодействие терроризму; машинное обучение; текстовая аналитика; тематическое моделирование; неотрицательная матричная факторизация.

Igor Mashechkin, Mikhail Petrovskiy, Irina Pospelova, Dmitry Tsarev

Lomonosov Moscow State University, Moscow, Russia

AUTOMATIC SUMMARIZATION AND KEYWORDS EXTRACTION METHODS FOR DISCOVERING EXTREMIST INFORMATION ON THE INTERNET

ANNOTATION

Nowadays there are growing number and damage of terrorist attacks carried out by lone terrorists under the influence of propaganda and extremist ideologies and by organized terrorist community, having a weakly connected network structure. The primary tool for information exchange, recruitment and propaganda for such structures is the Internet, namely web resources, social networks and e-mail. In this connection there is the problem of detecting, communication topics exrtraction, mining relationships, further monitoring and predicting threats from individuals, groups and network communities, generating and distributing terrorist and extremist information on the Internet. The present work is devoted to the research and development of machine learning methods aimed at discovering potentially dangerous information on the Internet. A new method proposed for automatic summarization and keywords extraction to discover extremist content in the flow of text messages. The applicability and effectiveness of the proposed method is experimentally demonstrated on the benchmark dataset collected in the framework of Dark Web project.

KEYWORDS

Security and counter-terrorism; machine learning; text analysis; topic model; non-negative matrix factorization.

* **Труды I Международной научной конференции «Конвергентные когнитивно-информационные технологии» (Convergent'2016), Москва, 25-26 ноября, 2016**

Введение

За последнее десятилетие террористические и экстремистские организации значительно увеличили свое присутствие в сети Интернет и социальных сетях, активно используя эти средства для вербовки новых членов и их обучения, подготовки и организации террористических атак, пропаганды насилия, распространения экстремисткой литературы и т.п. Использование Интернет – свободного и открытого ресурса – позволяет быстро и анонимно распространять любую информацию, обращаться напрямую к аудитории социальных сетей и форумов, не опасаясь цензуры, присутствующей в традиционных средствах массовой информации. Мероприятия, направленные на выявление террористов и связанных с ними лиц, пресечение распространения экстремистских материалов, предотвращение готовящихся терактов требуют анализа всей информации, поступающей от представителей экстремистских группировок. В этом контексте анализ Интернет-ресурсов выходит на первый план. В силу огромного объема распространяемой через Интернет информации, ее языкового многообразия и требования ее мониторинга в режиме реального времени необходимо использовать автоматические процедуры текстового анализа с целью выявления потенциально опасных пользователей, своевременного удаления экстремистских материалов, анализа информации о террористах и готовящихся терактах. Основными задачами при создании автоматических средств анализа информации террористической направленности являются выбор подходящих данных для тестирования алгоритмов и разработка алгоритмов, пригодных для решения задачи выявления террористической активности.

Согласно исследованию [1], проведенному в 2015 году, использование социальных сетей для отслеживания распространения радикальных идей и экстремистских угроз привлекает внимание исследователей уже более 10 лет. В последние 3 года наблюдается всплеск исследовательского интереса к идентификации и прогнозированию на основе анализа текстового содержания сообщений в открытых социальных сетях. Авторы [1] отмечают, что наиболее часто в качестве источника данных выступает Twitter, а для анализа содержания используются различные методы поиска информации и машинного обучения. Кластеризация, логистическая регрессия и динамическое расширение запроса (Dynamic Query Expansion) больше подходят для прогнозирования террористических актов, беспорядков или протестов. Общей компонентой различных подходов и методов является распознавание именованных сущностей (Named Entity Recognition, NER), позволяющее извлекать структурированную информацию из неструктурированных или слабоструктурированных документов. Для выявления радикализма и экстремизма в режиме реального времени чаще всего используются метод k-ближайших соседей (K Nearest Neighbor), наивный байесовский классификатор (Naive Bayes), метод опорных векторов (Support Vector Machine, SVM), деревья решений, Topical Crawler/Link Analysis и другие.

В работах, основанных на анализе общедоступной информации в Интернете (Twitter, текстовые документы свободного доступа), одной из основных задач является выявление террористических сайтов и сообщений террористов. Трудность состоит в том, что, во-первых, общение на форумах осуществляется на разных языках, а также, возможно, и в их комбинации (это же касается размещаемых в Интернете документов). А во-вторых, в том, что простой поиск по ключевым словам или конкретным фразам не позволяет отличить террористические сайты от, например, сообщений новостных агентств. Кроме того, террористические сайты зачастую маскируются под новостные сайты и религиозные форумы. Число сайтов огромно, что делает их анализ в ручном режиме неэффективным, поэтому для корректной идентификации настоящих сайтов и форумов, связанных с определенными террористическими группами, необходимы автоматические средства эффективного отбора и фильтрации. Более сложной является задача определения принадлежности распространяемой информации к одной из террористических групп, поскольку разные террористические группы могут быть идеологически близкими и использовать схожую лексику.

В работе [2] предложено использовать деревья решений для классификации текстов, представленных в виде графов. Полученные в результате анализа документов подграфы позволяют выделить несколько слов, наличие которых в тексте однозначно определяет его принадлежность к террористическому сайту. В то же время, отсутствие всех этих слов означает, что документ точно не является террористическим.

Близкая задача, для решения которой используются несколько другие подходы, рассматривается в работе [3]. Здесь делается попытка автоматического определения радикального содержания, выпущенного джихадистскими группами в Twitter. Для этого сравниваются результаты классификации твитов на радикальные и нерадикальные с помощью методов SVM с линейной kernel –функцией, AdaBoost и наивный байесовский классификатор.

В [4] задача выявления твитов, пропагандирующих ненависть и экстремизм, решается как задача бинарной классификации с помощью методов k-ближайших соседей и LIBSVM. Показано, что классификация с помощью LIBSVM является более точной.

Другое направление исследования экстремистских текстов в Интернете состоит в определении типа активности интернет пользователей. В работе [5] по данным записей в Twitter решается задача выявления пользователей-экстремистов, а также оценивается, будет ли обычный пользователь выбирать экстеримистские материалы и будут ли пользователи отвечать на контакты, инициированные экстермистами. При этом анализ может выполняться на агрегированных данных постфактум либо в режиме прогноза в реальном времени.

В работе [6] представленая система The Advanced Terrorist Detection System (ATDS), предназначенная для отслеживания в режиме реального времени доступа к аномальному контенту, который может включать в себя созданные террористами сайты, путем анализа содержания информации, полученной пользователями через Интернет. ATDS фунционирует в режиме обучения и распознавания. В режиме обучения ATDS определяет типичные интересы заранее определенной группы пользователей путем обработки web-страниц, к которым эти пользователи обращались в течении некоторого времени. В режиме распознавания ATDS осуществляет в реальном времени мониторинг интернет-траффика, создаваемого контролируемой группой, анализирует содержание web-страниц и сигнализирует, если полученная информация не входит в типичный круг интересов группы и является схожей с интересами террористов. Система анализирует произвольные текстовые данные, по которым с помощью кластеризации по методу k средних определяются типичные интересы пользователей (групп пользователей).

В работе [7] ставится задача выявления шаблона активности, типичного для террористов. Кластеризация исполнителей терактов по схожести дает такую значимую информацию, как общие характеристики различных групп, типичные цели терактов и используемое оружие. В данном исследовании разработан метод классификации террористических групп по примерам их атак, основанный на анализе текстуального описания этих атак с использованием латентного семантического индексирования и кластеризации. В качестве источника исходных данных использовался START (Study of Terrorism and Responses to Terrorism) с 1970 по 2010 годы [8].

Все вышеперечисленные исследования основаны на решении задач классификации и категоризации в случае, когда, как правило, есть предположения относительно тематик анализируемых (интересующих) текстовых документов. Однако для более глубокого тематического анализа текстов, необходимого, например, для систематизации сведений о террористической и экстремистской активности, идентификации типа активности, исследования эволюции террористических групп, требуются иные подходы.

Решение задачи тематического анализа осложняется рядом факторов. Информация, распространяемая террористическими группами, разнородна, сообщения в социальных сетях достаточно короткие, содержат сленг и закодированные слова, что делает бессмысленным семантический анализ. Наиболее часто в такой ситуации используется метод скрытого распределения Дирихле (Latent Dirichlet Allocation, LDA) [9].

В англоязычной литературе в последнее время появилось довольно много работ, в которых тестирование алгоритмов анализа текстов экстремистского содержания проводится на данных Dark Web. Эти данные были собраны учеными Аризонского университета (The University of Arizona) с различных форумов и сайтов выявленных террористических организаций [10, 11]. Появление Dark Web дало импульс к проведению большого числа разнообразных исследований, основанных на тематическом анализе его данных, позволяющих решать гораздо более сложные задачи, нежели бинарная классификация.

Работы [12, 13] посвящены решению важной для антитеррористических приложений задачи раскрытия подгрупп пользователей, чьи основные предметы обсуждения могут представлять угрозу национальной безопасности. Сложность состоит в том, что большинство алгоритмов выявляют разделенные сообщества, это значит, что каждый член сообщества принадлежит только к одному сообществу. Таким образом, часть информации о членах сообщества игнорируется, что приводит к неверной интерпретации результатов выявления групп. В данной работе предлагается подход, комбинирующий традиционные методы сетевого анализа для выявления перекрывающихся сообществ со средствами текстового анализа тематических моделей. Затем разрабатывается алгоритм определения подгрупп (под-сообществ). Для выявления тематик в работе применяется LDA, который в комбинации с алгоритмом all-previous-reply позволяет построить сеть взаимосвязей участников форума по набору тематик.

Работа [14] исследуют возможность идентификации вербовочной активности агрессивных групп на экстремистских сайтах социальных сетей. В другой работе этих авторов [15] представлено

исследование прогнозирования уровня ежедневной активности кибер-вербовки агрессивных экстремистских групп. Для идентификации вербовочных постов используется модель на основе SVM. Текстовое содержание анализируется с помощью LDA. Результаты анализа подаются в различные модели временных рядов для прогнозирования активности вербовки. Количественный анализ показывает, что использование основанных на LDA тематик в качестве предикторов в моделях временных рядов уменьшает ошибку прогнозировния по сравнению со сучайным блужданием, авторегрессией проинтегрированного скользящего среднего и экспоненциальным сглаживанием.

Схожий подход предлагается в работе [16], посвященной решению задачи выявления ключевых членов сообщества на основе тематик, для чего комбинируются инструменты интеллектуального анализа текстов и анализа социальных сетей. Сначала с помощью LDA по данным форума строятся две основанные на тематике сети: первая ориентирована на точку зрения создателя темы, а вторая – на отвечающих всего фоума. Затем с помощью различных средств сетевого анализа выделяются ключевые члены обсуждения тематики. Эксперименты успешно проведены на англоязычных форумах, доступных в Dark Web.

В [17] предлагается подход для раскрытия скрытых тематик в содержании сайтов экстремистской напрвленности. Содержание и данные сайтов (в данной работе www.natall.com) собираются поисковым роботом и экспортируются в документы. Для анализа выделенных документов с целью отыскания скрытых тематик на сайтах террористов и экстремистов используется LDA.

Как видно, развитие подходов к представлению текстовой информации, ее обработке, построение эффективных и точных алгоритмов анализа текстов, выявления их тематик является важным и актуальным научным направлением, которому в мире уделяется большое внимание. Следует отметить, что русскоязычные публикации, посвященные анализу информации террористической направленности с помощью математических методов, практически отсутствуют. По-видимому, это связано и с нехваткой систематизированных данных для тестирования алгоритмов, и с отсутствием выраженной потребности в автоматической обработке и поиске информации в Интернете (поскольку такая обработка осуществляется вручную экспертами).

Таким образом, разработка автоматических средств тематического анализа позволит существенно повысить эффективность решения задач поиска в Интернете документов и отдельных сообщений террористической и экстремистской направленности, что, в свою очередь, приведет к возможности предотвращения готовящихся терактов, уменьшению влияния экстремистских групп и повышению уровня национальной безопасности.

Предалагемый метод автоматического аннотирования и выделения ключевых слов

В качестве основных типов источников информации в Интернет обычно рассматриваются сообщения социальных сетей для публичного обмена сообщениями (таких как Twitter), публикации блогов, форумов и электронных СМИ. Анализируемые текстовые сообщения и документы можно представить в виде совокупности самого текста сообщения и характеризующих его набора атрибутов. В общем случае текст сообщения имеет произвольный объём, т.е. может быть как коротким текстовым сообщение, так и большим текстовым документом, в том числе и лентой коротких текстовых сообщений. Обязательным атрибутом сообщения является временная метка регистрации сообщения, кроме того, возможны дополнительные атрибуты, например, отправитель и получатель, которые могут использоваться для построения топологии группы пользователей или сетевого сообщества. Сообщения могут быть чрезмерно короткими, состоящими из нескольких слов, или слишком большими текстовыми документами. В первом случае возникает задача объединения нескольких близких по времени сообщений одного отправителя в одно общее сообщение (ветку обсуждения). Во втором случае, наоборот, решается задача сокращения объема сообщения с сохранением большей части информации, т.е. задача автоматического аннотирования – выделения ключевых фрагментов документа. Причем зачастую приходится решать обе эти задачи последовательно: сначала объединять сообщения в ленту или ветку обсуждения, а потом автоматически аннотировать или реферировать ее, выделяя наиболее типичные или статистически значимые сообщения. Кроме того, важным фактором является язык написания сообщения. Помимо традиционной проблемы использования различных, в том числе восточных и ближневосточных языков, в ресурсах террористического и экстремистского содержания может использовать сленг или жаргон, употребляемый только в узком кругу пользователей. Также могут использоваться специальные кодовые слова для замены ключевых слов, по которым обычно осуществляется поиск, таким как названия наркотиков, оружия, имена конкретных лиц и названия географических мест. Все эти особенности делают крайне тяжелым применение традиционных

методов NLP в обозначенных задачах, поэтому в настоящей работе предлагается сделать акцент на языково-независимые методы анализа текстов, преимущественно статистические с выделением признаков текстов на основе базовых словоформ и латентно-семантического анализа.

Для решения задачи автоматического аннотирования необходимо сформировать набор наиболее значимых фрагментов исходного текста. На сегодняшний день наиболее популярные методы автоматического аннотирования, которые вычисляют релевантность фрагментов текста, основаны на тематическом моделировании текстов с использованием латентно-семантического анализа [18, 19]. Латентно-семантический анализ работает с матричным представлением коллекции текстов, получаемым с помощью модели «мешок слов» (англ. «bag-of-words») [20].

В задаче автоматического аннотирования в качестве текстов используются отдельные фрагменты документа, например, предложения. Каждый фрагмент j ($1 \le j \le n$) представляется в виде числового вектора $A_j = [a_{1,j}, a_{2,j}, \ldots, a_{m,j}]^T$ фиксированной размерности m, где m — число признаков коллекции фрагментов, а i-я ($1 \le i \le m$) компонента вектора A_j определяет вес i-го признака в j-ом фрагменте. Таким образом, коллекция фрагментов документа представляется в виде числовой матрицы $A \in \mathbb{R}^{m \times n}$, строки которой соответствуют признакам, а столбцы — фрагментам. В качестве признаков в модели «мешок слов» используются термы — лексемы, входящие в текст. Однако обычно применяются некоторые меры по предварительной обработке лексем текста для получения более «информативного» признакового пространства: удаление стоп-слов, приведение слов к нормализованной форме (стемминг) и т.д. Цель предварительной обработки текста — оставить только те признаки, которые наиболее информативны, т.е. наиболее сильно характеризуют текст. К тому же сокращение числа анализируемых признаков приводит к уменьшению объема используемых вычислительных ресурсов.

Следующим шагом латентно-семантического анализа является определение основных тематик документа и представление фрагментов текста в пространстве тематик. Для этого к матрице A применяется одно из матричных разложений, например, сингулярное разложение (англ. *Singular Value Decomposition*, SVD) и неотрицательная матричная факторизация (англ. *Non-negative Matrix Factorization*, NMF) [18, 19]. Предлагаемый метод вычисления релевантности фрагментов текста основан на оценке весов тематик в нормализированном пространстве тематик, получаемом с помощью неотрицательной матричной факторизации. Цель неотрицательной матричной факторизации, применённой к матрице $A \in \mathbb{R}^{m \times n}$, состоит в нахождении матриц $W_k \in \mathbb{R}^{m \times k}$ и $H_k \in \mathbb{R}^{k \times n}$ с неотрицательными элементами, которые минимизируют целевую функцию

$$f(W_k, H_k) = \frac{1}{2}\|A - W_k H_k\|_F^2, \; k << min(m, n). \quad (1)$$

Матрица W_k задает отображение пространства тематик размерности k в пространство термов размерности m, матрица H_k соответствует представлению текстов в пространстве тематик. Элементы матрицы W_k неотрицательны, поэтому можно установить, какие термы текста лучше всего характеризуют каждую из выделенных тематик, которым соответствуют столбцы матрицы W_k. Аналогично можно установить, какие из выделенных тематик наилучшим образом характеризуют каждый фрагмент. Данное свойство широко используется при кластеризации текстовых данных, где наиболее характерная тематика документа соответствует его кластеру. Таким образом, благодаря неотрицательности элементов матриц W_k и H_k, неотрицательная матричная факторизация, в отличие от сингулярного разложения, имеет хорошо интерпретируемое тематическое пространство.

После применения неотрицательной матричной факторизации к матрице текстовых фрагментов A выполняется нормировка пространства k тематик, т.е. приведение длин вектор-столбцов матрицы W_k к единице. Это необходимо, поскольку неотрицательная матричная факторизация даёт неединственное решение задачи (1). Если матрицы W_k и H_k являются решением (1), то матрицы $W_k \cdot D$ и $D^{-1} \cdot H_k$, где матрица D — любая положительная диагональная матрица размерности $k \times k$, также будут решением (1). Таким образом, используя разные значения диагональных элементов в D, можно получать преобладание различных тематик в тематическом представлении фрагментов H_k. Для решения проблемы корректной оценки весов получаемых тематик во фрагментах нормировка матрицы W_k производится следующим образом:

$$A_k = W_k H_k = Wn_k \cdot Hn_k, \quad (2)$$

где $Wn_k = W_k \cdot diag\left(\left\|w^1\right\|^{-1}, \ldots, \left\|w^k\right\|^{-1}\right)$, $Hn_k = diag\left(\left\|w^1\right\|, \ldots, \left\|w^k\right\|\right) \cdot H_k$, $\left\|w^l\right\| = \sqrt{\sum_{p=1}^{m} w_{pl}^2}$, $1 \le l \le k$.

Столбцы матрицы $Hn_k=[hn_{ij}]$ соответствуют n фрагментам в нормированном пространстве k

тематик. Каждая из k строк Hn_k соответствует вектору, показывающему, насколько сильно представлена соответствующая тематика в каждом из n фрагментов. Тем самым, чем больше длина вектор-строки матрицы Hn_k, тем соответствующая тематика «больше» представлена во всем документе. Исходя из этого, вес тематики l оценивается как длина l-й вектор-строки матрицы Hn_k:

$$\|hn_l\| = \sqrt{\sum_{q=1}^{n} hn_{lq}^2} = \|w^l\| \cdot \sqrt{\sum_{q=1}^{n} h_{lq}^2} = \|w^l\| \cdot \|h_l\|, \; 1 \le l \le k. \tag{3}$$

Тогда релевантность j-го фрагмента вычисляется как норма вектора, являющегося результатом поэлементного умножения вектора глобальных весов тематик и вектора весов тематик в рассматриваемом фрагменте:

$$R_j = \sum_{i=1}^{k} \left(\left(\|w^i\| \cdot \|h_i\| \right) \cdot \left(\|w^i\| \cdot h_{ij} \right) \right) = \sum_{i=1}^{k} \left(\|w^i\|^2 \cdot \|h_i\| \cdot h_{ij} \right). \tag{4}$$

Для составления аннотации выбирается некоторое число предложений с наибольшими значениями полученной релевантности. Таким образом, идея предложенного метода автоматического аннотирования заключается в выделении основных тематик в тексте документа и нахождении фрагментов текста, которые наилучшим образом описывают выделенные тематики, путём расчёта их релевантности. Выделенные тематики также можно описывать и набором ключевых слов, в силу описанного выше свойства интерпретируемости тематического пространства, формируемого с помощью неотрицательной матричной факторизации. Матрица W_k является представлением выделенных тематик в пространстве термов исходного текста, получаемом с помощью модели «мешок слов», каждый столбец данной матрицы соответствует отдельной тематики. Тогда для каждой тематики l ($1 \le l \le k$) выбирается p термов из словаря модели «мешок слов» с индексами (i_1, i_2, …, i_p), соответствующими максимальным элементам в l-м столбце матрицы W_k:

$$\{i_1, i_2, \ldots, i_p\} \mid \forall l \, \forall z \notin \{i_1, i_2, \ldots, i_p\} : W_{z,l} \le W_{i,l}, \; i \in \{i_1, i_2, \ldots, i_p\} \tag{5}$$

Релевантность фрагмента текста показывает его информационную значимость в рассматриваемом документе, поэтому релевантность можно рассматривать в качестве оценки количества информации, содержащейся в данном фрагменте. Исходя из этого, можно определить минимальное число фрагментов текста, требующееся для покрытия заданного процента содержащейся в тексте информации. Для построения результирующего документа, не содержащего информационного шума, выбираются его фрагменты с максимальными релевантностями, сумма которых не превышает заданный процент информации, как правило, равный не более 30% [21].

Экспериментальное исследование предложенного метода

Для подтверждения работоспособности предложенного метода автоматического аннотирования и выделения ключевых слов был проведен следующий эксперимент. Был взят набор эталонных данных под название “kavkazchat”, подготовленный в рамках проекта Dark web [10] в лаборатории Искусственного интеллекта университете Аризонского университета (University of Arizona), США. Этот набор данных содержит информацию, собранную на форумах, преимущественно посвященных проблемам и жизни российского Северного Кавказа, где в рамках проекта Dark Web были выявлены сообщения потенциально экстремистского и террористического содержания. Объем текстовых данных достаточно велик, весь набор содержит более 600 гигабайт текстовых данных, включая сообщения на русском языке в кириллице, на русском языке в транслите, на арабском языке, на национальных языках Северного Кавказа в кириллической транскрипции, а также опечатки и специальное написание слов, например, включение в слова цифры: «муджа1хид» («моджахед»).

В наборе данных содержится 16 тысяч веток обсуждения разной тематической направленности, в которых участвуют несколько тысяч пользователей. Объем веток обсуждения варьируется от одного килобайта и меньше до 5 мегабайт. Далеко не все ветки содержат информацию потенциально экстремистского содержания. Много сообщений посвящено обсуждению религиозных тем, таких как правила поведения в исламском обществе, взаимоотношения между мужчинами и женщинами в нем и т.п. Также присутствуют бытовые темы, такие как кулинария, обсуждение автомобилей и спорта. Много сообщений посвящено обсуждению политических событий в мире, так или иначе связанных с Россией, Кавказом и Ближним востоком, например, война в Афганистане и Ливии, события в Грузии, Польше, авария на атомной электростанции в Японии. Следует отметить, что простой «ручной» поиск по ключевым словам для такого типа данных дает крайне низкую точность. Например, в ветке, полностью посвященной

кулинарии и не содержащей экстремистской информации, могут быть комментарии вида «это очень полезно и питательно, поэтому подойдет моджахедам». В ветках, посвященных обсуждению политических событий, также используется близкая лексика, при этом зачастую грань между обычным комментарием и потенциально экстремистским может быть очень тонкая. Например, к вполне нейтральному новостному описанию события в горячей точке может быть добавлен комментарий, использующий словосочетание «русские оккупанты» или «американские террористы», что делает ветку подозрительной с точки зрения потенциального содержания экстремистской информации. Таким образом, выбранный набор данных является крайне интересным с точки зрения решения задачи тематического анализа, поскольку позволяет оценить качество предложенного подхода на существенно неоднородных данных.

В результате применения стандартного способа выявления скрытых тематик с использованием метода латентно-семантического анализа на основе сингулярного разложения [20] было выявлено 15 тематик, общие характеристики которых представлены в таблице 1.

Таблица 1.

N	Ключевые слова тематики	Комментарий
1	Аллах,ибн, пророк, посланник	Сообщения религиозной тематики
2	Ма,ца,хьо,ду,ю	Сообщения не на русском языке (содержательный анализ не проводился)
3	Народ, Россия, война, мусульманин	Сообщения на тему политической жизни в РФ
4	Аллах,говорить,знать,мусульманин	Сообщения религиозной тематики
5	Сердце, душа, глаз, любовь	Сообщения религиозной тематики
6	Россия,русский, Путин, Москва, страна	Сообщения на тему политической жизни в РФ
7	уа,est,eto,ti,je	Сообщения не на русском языке (содержательный анализ не проводился)
8	Сообщать,моджахед,США,военный,район	Обсуждение военных действий в мире
9	Ду,ца,ма,хьо,иза	Сообщения не на русском языке (содержательный анализ не проводился)
10	Район,моджахед,сообщать,Дагестан	Обсуждение военных действий в рамках контр-террористической операции
11	Чеченец, район, народ,Чечня	Сообщения на тему политической и общественной жизни в Чечне
12	Автомобиль,модель, компания, двигатель	Обсуждение автомобилей
13	Масло,вода,рецепт,организм,ложка	Обсуждение кулинарии
14	Сайт,русский,файл,скачать,программа	Обсуждение онлайн ресурсов в сети Интернет
15	Аллах,Кавказ,Грузия,война,Россия	Обсуждение военных действий в рамках войны с Грузией 2008 года

Применение алгоритма иерархической кластеризации [20] позволило сформировать 10 кластеров веток, общая характеристика получившихся кластеров представлена ниже в таблице 2.

Таблица 2.

N	Ключевые слова кластера	Процент веток в кластере
1	Аллах лучше например необходимо следует достаточно знаю брат некоторые потом	0.093932
2	Народ Россия война вообще думаю Путин русские чеченский Чечня ФСБ	0.105243
3	Аллах агентство солдат kavkazcenter ummanews безопасность город источник	0.07937
4	Банда муртад район ФСБ Дагестан города слова сообщает отдел	0.057121
5	Аллах брат говорит дом думаю знаю интересно Ислам	0.255359
6	США Ислам мир мусульманин ссылка заявил ummanews пишет страна сми слова	0.020686
7	Аллах дал чеченский Чечня брат район говорят думаю война	0.010687
8	Агентство вопрос дело отдел Россия kavkazcenter банда безопасность власти война заявил	0.233235
9	США военные передает безопасность агентство территория сми город ссылка	0.039498
10	Дагестан отдел убит kavkazcenter банда город источник кавказ-центр ссылка кафир	0.104868

Анализ набора данных с помощью модели кластеризации, в отличие от тематической модели, в которой тематики ортогональны, т.е. не коррелируют друг с другом, показывает, что в большинстве веток обсуждений присутствует политическая составляющая, в том числе (во многих кластерах) потенциально содержащая экстремистскую информацию.

Применим разработанный метод автоматического построения аннотаций и выделения ключевых слов, описанный в предыдущем разделе. Он позволит сформировать из исходного набора данных несколько наборов, содержащих вместо исходных веток сообщений аннотации, соответствующие заданной доли сохраненной информации, или ключевые слова. В настоящей работе метод использовался со следующими настройками: число «внутренних тематик» k из формулы (1) в рамках каждой ветки полагалось равным $k=3$. Процент сохраненной информации (суммарная доля релевантностей фрагментов текста по формуле (4)) выбиралась 30% и 10%,

соответственно. Для каждой ветки был сформирован набор из 15 ключевых слов. Таким образом, мы получили три набора данных, названных

- NMF30 – набор аннотаций исходного набора, в котором каждая аннотация суммарно содержит предложения, покрывающие 30% от общей релевантности всех предложений, полученный объем аннотаций в диапазоне от 1 до 300 килобайт, в среднем 3 килобайта;
- NMF10 - набор аннотаций исходного набора, в котором каждая аннотация суммарно содержит предложения, покрывающие 10% от общей релевантности всех предложений, полученный объем аннотаций от 1 до 200 килобайт, в среднем 2 килобайта;
- KWORDS – набор ключевых слов по каждой из веток (около 30 слов для каждой ветки).

В таблице 3 приведены примеры построенных аннотации (наиболее релевантных ветке предложений) и ключевых слов для некоторых отобранных веток, наиболее разных по тематикам веток.

Таблица 3.

Основная тема ветки	**Наиболее релевантное предложение из аннотации**	**20 ключевых слов по всей ветке**
Потенциальный экстремизм (относится к тематике «Обсуждение военных действий в рамках контр-террористической операции»)	«Статистика Джихада в Имарате Кавказ за месяц Мухаррам 1432 года по Хиджре (Декабрь 2010) Вилаят Нохчийчоь: Количество всех проведённых моджахедами операций – 5 Убито муртадов – 1 Ранено муртадов – 3 Убито кафиров – 1 Ранено кафиров – 2 Вилаят Гlалгlайче: Количество всех проведённых моджахедами операций – 9 Количество крупных проведённых моджахедами операций – 2: 1) Ликвидация командира бандгруппы ОВД по Назрановскому району 2) Ликвидация заместителя начальника штаба войсковой части кафира Александра Орлова Убито муртадов – 1 Ранено муртадов – 2 Убито кафиров – 1 Ранено кафиров – 1 Моджахедов стало Шахидами (иншаАллах) – 4»	район сообщать сотрудник орган правоохранительный мвд республика боевик Дагестан источник муртад кафир моджахед бой вилайят ранен убит банда http
Бытовые темы без экстремизма (тематика «Кулинария»)	«ты давала, только ты другое сделала рис, кукуруза и колбаса что ли))) ну у меня по той видео токо формачка сделана так) а остальное всё по своему сделала) в салате у меня - рис, мелко нарезанный лук, пожаренная колбоса кубиками, солёные огурцы, кукуруза, укропа много, и майонез=) а блинчики сделала по своему) там у неё они очень тонкие а у меня средние))»	тесто яйца мука масло сахар соль добавлять ложка духовка начинка готов мясо блюдо салат курица ел соус картошка вчера вкусно

Даже по приведенным выше фрагментам понятна основная идея разработанного метода, который позволяет по достаточно длинному тексту ветки сообщений найти адекватные ключевые слова и наиболее релевантные предложения, описывающие основную суть ветки обсуждения. Теперь проверим формально, что построенные с помощью предложенного метода наборы аннотаций и ключевых слов, незначительно теряют или вообще не теряют ключевую информацию, содержащуюся в исходных текстах. Для этого применим построенные для исходного набора данных тематическую и кластерную модели к полученным сокращенным наборам аннотаций и ключевых слов NMF30, NMF10, KWORDS. Если в результате сокращения значимая информация не была потеряна, то аннотации и документы с ключевыми словами попадут в те же тематики и кластеры, что и исходные полные документы.

Результаты эксперимента для тематической модели приведены в таблице 4.

Таблица 4.

Набор данных	**Оценка точности по тематикам**														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
KWORDS	0.86	0.86	0.87	0.84	0.86	0.87	0.86	0.87	0.92	0.88	0.87	0.87	0.86	0.85	0.87
NMF10	0.9	0.88	0.88	0.86	0.87	0.86	0.88	0.87	0.93	0.86	0.89	0.88	0.87	0.87	0.92
NMF30	0.92	0.9	0.9	0.89	0.89	0.88	0.9	0.9	0.94	0.9	0.91	0.9	0.9	0.89	0.93

Приведенные в таблице 4 результаты показывают, что все три полученных сокращенных набора имеют высокую степень согласованности с тематической моделью, построенной на несокращенном текстовом наборе. При этом, чем больше сохраняется информации в аннотации, тем выше согласованность, но даже набор ключевых слов, оставленный из исходного текста, дает согласованность на уровне 86-87% по всем тематикам.

В таблице 5 приведены результаты для кластерной модели.

Таблица 5.

Набор данных	Попадание аннотации и полного документа в один кластер
KWORDS	0.6143991001
NMF10	0.9238797575
NMF30	0.9285669646

Видно, что согласованность результатов кластеризации заметно ниже для набора ключевых слов, а для аннотаций остается на очень высоком уровне, более 92%, причем, аналогично тематической модели, чем больше сохраняется информации, тем выше согласованность. Распределение согласованности в зависимости от номера кластера представлено в таблицах 6-8 по каждому из сокращенных наборов.

Таблица 6

Набор NMF10											
	Кластер аннотации										
Кластер документа	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**20**	**Total**
1	1213	13	1	15	230	16	11	3	1	0	1503
2	3	1495	4	13	29	4	14	116	1	5	1684
3	4	0	1255	1	0	0	0	1	1	8	1270
4	3	17	21	779	14	1	4	12	5	58	914
5	184	82	4	11	3754	10	29	5	1	6	4086
6	8	21	6	4	9	226	4	46	7	0	331
7	2	9	15	17	11	2	100	5	1	9	171
8	0	15	2	3	0	4	1	3692	14	1	3732
9	1	1	17	3	0	1	0	11	597	1	632
10	0	0	0	5	0	0	0	0	1	1672	1678
Total	1418	1653	1325	851	4047	264	163	3891	629	1760	16001

Таблица 7.

Набор NMF30											
	Кластер аннотации										
Кластер документа	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**20**	**Total**
1	1245	13	2	12	197	19	12	1	2	0	1503
2	2	1506	4	15	25	3	11	112	1	5	1684
3	3	0	1257	1	0	0	0	1	1	7	1270
4	3	15	19	788	13	1	4	10	7	54	914
5	187	78	5	9	3769	9	22	3	1	3	4086
6	10	21	5	7	7	227	3	45	6	0	331
7	2	10	18	18	9	3	101	4	1	5	171
8	0	15	2	3	0	4	1	3693	13	1	3732
9	1	0	15	1	0	1	0	13	600	1	632
10	0	0	0	5	0	0	0	0	1	1672	1678
Total	1453	1658	1327	859	4020	267	154	3882	633	1748	16001

Из таблиц 6 и 7 видно, что при применении моделей аннотирования практически все кластеры распознаются хорошо, а при применении ключевых слов (см. таблицу 8) основные ошибки возникают при распознавании 6, 7 и 9 кластеров. Ошибки на этих кластерах объясняются, во-первых, их небольшим размером, а, во-вторых, тем, что их ключевые слова достаточно сильно пересекаются (см. таблицу 2 с описанием кластеров).

Таблица 8

Набор KWORDS											
	Кластер аннотации										
Кластер документа	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**20**	**Total**
1	705	29	10	89	594	21	26	8	18	3	1503
2	24	1017	51	93	163	19	33	185	31	68	1684
3	8	12	714	18	29	5	9	12	58	405	1270
4	27	60	66	338	60	6	17	31	31	278	914
5	331	91	37	79	3407	13	76	8	20	24	4086
6	69	60	23	21	66	22	10	42	17	1	331
7	9	10	28	25	25	2	28	3	12	29	171
8	54	729	151	146	79	93	40	1937	335	168	3732
9	17	19	180	42	18	7	13	92	201	43	632
10	10	7	46	126	14	0	5	3	5	1462	1678
Total	1254	2034	1306	977	4455	188	257	2321	728	2481	16001

Выводы

В настоящей работе рассматривается важная прикладная задача использования методов машинного обучения для выявления потенциальной экстремистской и террористической информации в сети Интернет. Дается обзор существующих решений и подходов и предлагается новый оригинальный метод автоматического аннотирования и выделения ключевых слов с удалением информационного шума, основанный на использовании неотрицательной матричной факторизации для матрицы термов веток текстовых сообщений из сети Интернет. Применимость и эффективность предложенного метода демонстрируется экспериментально на эталонном наборе реальных Интернет данных, потенциально содержащих информацию экстремистского характера. В эксперименте показано, что применение предложенного метода позволяет:

- получать содержательные и релевантные аннотации в виде выдержки наиболее важных предложений из исходного текста;
- генерировать по тексту релевантные ключевые слова, которые отражают основную суть исходного текста и, кроме того, могут быть впоследствии использованы для поиска соответствующей информации в сети Интернет;
- значительно сократить объемы анализируемой информации при незначительной потере точности тематических и кластерных моделей, которые были построены для несокращенных текстов набора.

В дальнейшем предполагается продолжить исследования в этом направлении и решить задачи:

- языково-независимого аннотирования и генерации ключевых слов с учетом смеси различных языков с использованием подхода на основе n-грамм;
- разработать признаковое пространство для текстовых сообщений в сети Интернет, включающее языково-независимые тематические признаки, информацию о ссылках и внешних Интернет ресурсах, упоминаемых в сообщении, хэштегах и информацию об авторах сообщений;
- реализовать системы непрерывного мониторинга, аннотирования и тематического моделирования потоков текстовых сообщений, лент, записей в форумах и социальных сетях интернет сообществ с целью непрерывного поиска и выявления потенциально экстремистской информации.

Работа выполнена при финансовой поддержке гранта РФФИ № 16-29-09555\16 по направлению «Безопасность и противодействие терроризму».

Литература

1. Swati Agarwal, Ashish Sureka Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats arXiv:1511.06858 [cs.CY].

2. Last, Mark, Markov, Alex, Kandel, Abraham, Chen, Hsinchun, Yang, Christopher C. Multi-lingual Detection of Web Terrorist Content. Intelligence and Security Informatics: Techniques and Applications, 2008, Springer Berlin Heidelberg, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-540-69209-6_5 P 79-96.
3. Enghin Omer Using machine learning to identify jihadist messages on Twitter http://uu.diva-portal.org/smash/get/diva2:846343/FULLTEXT01.pdf.
4. Ashish Sureka; Swati Agarwal Learning to Classify Hate and Extremism Promoting Tweets Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint Year: 2014 Pages: 320 - 320, DOI: 10.1109/JISIC.2014.65.
5. Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, Aram Galstyan (2016) Predicting online extremism, content adopters, and interaction reciprocity arXiv:1605.00659 [cs.SI].
6. Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M. and Kandel, A. (2010), Detection of access to terror-related Web sites using an Advanced Terror Detection System (ATDS). J. Am. Soc. Inf. Sci., 61: 405–418. doi:10.1002/asi.21249.
7. Ibrahim Toure; Aryya Gangopadhyay Analyzing_terror_attacks_using_latent_semantic_indexing , 2013 IEEE International Conference on Technologies for Homeland Security (HST) Year: 2013 Pages: 334 - 337, DOI: 10.1109/THS.2013.6699024
8. http://www.start.umd.edu/start/.
9. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research. **3** (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
10. Yulei Zhang, Shuo Zeng, Li Fan, Yan Dang, Catherine A. Larson, and Hsinchun Chen. 2009. Dark web forums portal: searching and analyzing Jihadist forums. In Proceedings of the 2009 IEEE international conference on Intelligence and security informatics (ISI'09). IEEE Press, Piscataway, NJ, USA, 71-76.
11. Ahmed Abbasi and Hsinchun Chen Applying authorship analysis to extremist-group web forum messages, IEEE Intelligent Systems, 2005, V.20, pp. 67—75.
12. Sebastián A. Ríos and Ricardo Muñoz. 2012. Dark Web portal overlapping community detection based on topic models. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics* **(**ISI-KDD '12). ACM, New York, NY, USA, , Article 2 , 7 pages. DOI=http://dx.doi.org/10.1145/2331791.2331793.
13. Tope Omitola, Sebastián A. Ríos, John G. Breslin: Social Semantic Web Mining. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers 2015.
14. J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," Security Informatics, vol. 3, no. 1, pp. 1–10, 2014. doi:10.1186/s13388-014-0005-5.
15. Jacob R. Scanlon, Matthew S. Gerber: Forecasting Violent Extremist Cyber Recruitment. IEEE Trans. Information Forensics and Security 10(11): 2461-2470 (2015).
16. Gaston L'Huillier, Hector Alvarez, Sebastián A. Ríos, and Felipe Aguilera. 2011. Topic-based social network analysis for virtual communities of interests in the dark web. *SIGKDD Explor. Newsl.* 12, 2 (March 2011), 66-73. DOI=http://dx.doi.org/10.1145/1964897.1964917.
17. Li Yang and Feiqiong Liu and Joseph Migga Kizza and Raimund K. Ege Discovering Topics from Dark Websites IEEE Symposium on Computational Intelligence in Cyber Security, 2009. CICS '09, pp. 175 - 179, DOI: 10.1109/CICYBS.2009.4925106.
18. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V., Popov D.S. Automatic text summarization using latent semantic analysis // Programming and Computer Software. – 2011. – Т. 37. – №. 6. – С. 299-305.
19. Машечкин И. В., Петровский М. И., Царёв Д. В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование. – 2013. – Т. 14. – №. 1. – С. 91-102.
20. Manning C. D. et al. Introduction to information retrieval. – Cambridge: Cambridge university press, 2008. – Т. 1. – С. 496.
21. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V. Using NMF-based text summarization to improve supervised and unsupervised classification // Hybrid Intelligent Systems (HIS), 2011 11th International Conference on. – IEEE, 2011. – С. 185-189.

Поступила 21.10.2016

Об авторах:

Машечкин Игорь Валерьевич, д.ф.-м.н., профессор кафедры АСВК факультета ВМК МГУ, заведующий лабораторией Технологий программирования, mash@cs.msu.su**;**

Петровский Михаил Игоревич, к.ф.-м.н., доцент кафедры АСВК факультета ВМК МГУ, michael@cs.msu.su;

Поспелова Ирина Игоревна, к.ф.-м.н., доцент кафедры Исследования операция факультета ВМК МГУ, ipospelova05@yandex.ru**;**

Царёв Дмитрий Владимирович, мнс. лаборатории Технологий программирования факультета ВМК МГУ, tsarev@mlab.cs.msu.su**.**